Exploiting Document Image Analysis in the Humanities

Frédéric Rayar, Jean-Yves Ramel,, Rémi Jimenes

Exploiting Document Image Analysis in the Humanities

RAYAR Frédéric [(1)], RAMEL Jean-Yves [(1)], JIMENES Rémi [(2)]
Université François Rabelais Tours
[(1)] Laboratoire d'Informatique
[(2)] Centre d'Études Supérieures de la Renaissance
rayar@univ-tours.fr, ramel@univ-tours.fr, remi.jimenes@gmail.com

Introduction

With the growing popularity of the Internet and web-based technologies, digital libraries and archives are becoming increasingly interested in the mass digitization and transcription of ancient documents in order to render them available to a much wider class of users. The usefulness of transcribing the information they contain is well recognized. However, the software required to produce satisfactory automatic transcriptions is not currently available. Commercial Optical Character Recognition (OCR) systems are unsuited to deal with either the often impaired quality of ancient documents (worn by use, torn, stained, etc.) or the difficulties presented by non-standard fonts.

This context provided the motivation to set up, in collaboration with the Centre d'Études Supérieures de la Renaissance (CESR) of Tours [1], the Pattern Redundancy Analysis for Document Image Indexing and Transcription (PaRADIIT) research project [2]. Further details on this project can be found on the PaRADIIT website and in a previous publication [3].

Content Extraction and Typographic Studies

With the experience acquired in the PaRADIIT project, it became clear to us that our layout analysis software could easily be used to extract graphical contents such as initial letters, portraits, ornamental bands, etc. The website set up by the Bibliothèques Virtuelles Humanistes (BVH) team of the CESR [4] provides an example of how the innovative technical tools developed by our team have been and continue to be used to answer the specific requirements of current research in the humanities; in this case, by supplying the means of extracting and indexing many different types of graphical contents found in ancient French books [5].

This work led us to think about developing indexing tools which could be applied to texts, and, more specifically, to letters. Our aim here was to extract data from Early Modern books in order to create new font packages that could be used to further research in the history of printing materials. Relevant criteria were determined with the help of works by typography historians such as H. D. L. Vervliet [6a], [6b]. Figure 1 shows an XML schema of these criteria. Following this specification step, a Font Creation Tool was embedded in one of the project's software programs. A screenshot of this tool is shown in Figure 2. Sampling of Early Modern fonts and character styles is currently underway at the CESR.

```
<?xml version="1.0" encoding="utf-16"?>
<!--RETRO Model file-->
<Model>
 <Metadata>
  <Book Author="" Title="" Place="" PrinterOrPublisher="" Date="" Format="" />
  <Copy Library="" CallNumber="" Digitalization="" Copyright="" CataloguerName="" />
 </Metadata>
 <Transcription Character="" Unicode="" />
 <Image Filename="" Folder="" Page="" Resolution="" />
 <Thumbnail Name=□□ Width="" Height="" PositionX="" PositionY="" />
 <Typography IsSmallCap="□Type="" Alphabet="" Family="" SubFamily="" BodyHeight=""
        Thickness="" />
 <Description References="" Engraver="" Comments="" />
</Model>
```

Figure 1. XML Schema of a Font Indexing Criteria Model

From the outset, these new font packages appeared to us to offer various possibilities in terms of exploitation.

Our first idea, influenced by our OCR expertise, was to use them as input for Optical Font Recognition (OFR) systems, OCR mono-font systems, or OCR loop systems incorporating knowledge exploitation, in order to significantly improve the quality of transcriptions made from ancient books.

A second possibility for exploitation was rapidly pinpointed in the digital humanities field. The identification of typographic materials clearly has potential for specialists researching various aspects of printing, notably its aesthetic (i.e. the thickness and shapes of printing types) and economic history. Our font indexing tools may also shed new light on the book trade during the Renaissance by furthering research on the circulation of printing tools and materials. Up until the mid 16th century, printing types often passed from workshop to workshop, printers frequently selling or lending types to fellow printers. Afterwards, the printing materials trade came to be dominated by a small number of important type foundries based in Antwerp, Frankfurt and Paris. The identification of fonts used in the 17th and 18th centuries consequently promises to help establish the influence of each of these type foundries in Europe right up until the “graphic revolution” of the late 18th century (initiated by Baskerville and pursued by Bodoni, Fournier, and then Didot).

Inventory work also prepares the ground for the encoding and displaying of special characters and eventually their integration in Unicode Standard and Private Use Areas. This kind of work is already being carried out in the framework of projects like the Medieval Unicode Font Initiative (MUFI) [7], focusing on medieval manuscripts, and appears to be providing stimulus in typography research [8], [9].

Discussions with experts in literary fields may well provide other ideas about exploitation potential.

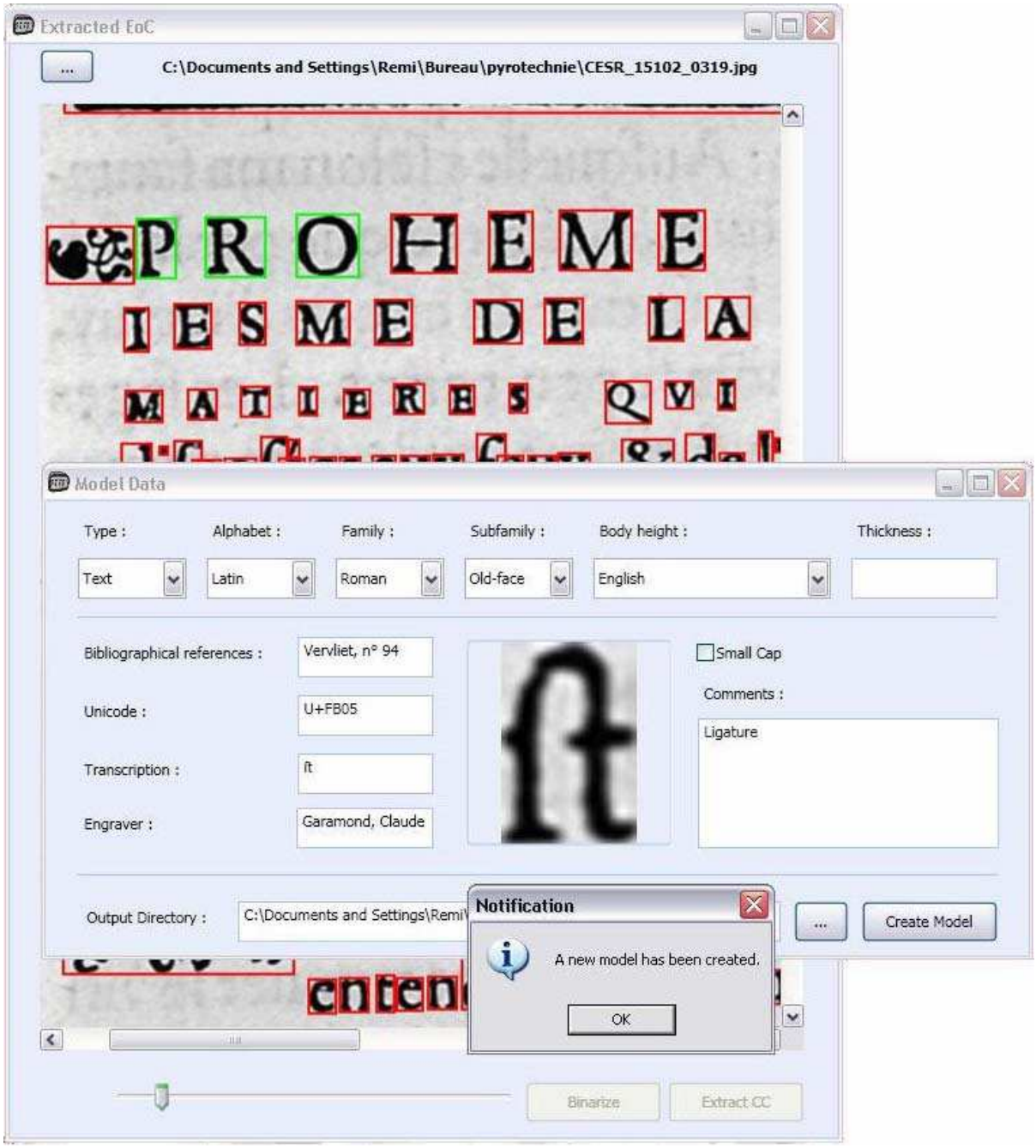

Figure 2. Screenshot of a Font Model Creation Tool

Towards an Online Collaborative Platform

Pursuing the idea of enabling the results of our work to further research in the humanities, we started to think about developing an online platform. This platform would not only enable users to consult digitized ancient books, but also to access the layout analysis and text transcriptions of these books. Simple search options using words and expressions taken from existing transcriptions have already been implemented. These could be complemented by alternative methods of searching contents, such as *word spotting* [10].

The Font Creation Tool presented above and others typography oriented tools, such as measurement of body height, could be embedded in this platform. The creation of an annotation environment, aimed at enabling experts to actively collaborate and participate in the creation and management of digital humanities resources, is currently being studied.

In order to make this platform as easy to use as possible, we have adopted a user-friendly web interface, with classic online library functions. Figure 3. shows a screenshot of our Book Explorer prototype [11].

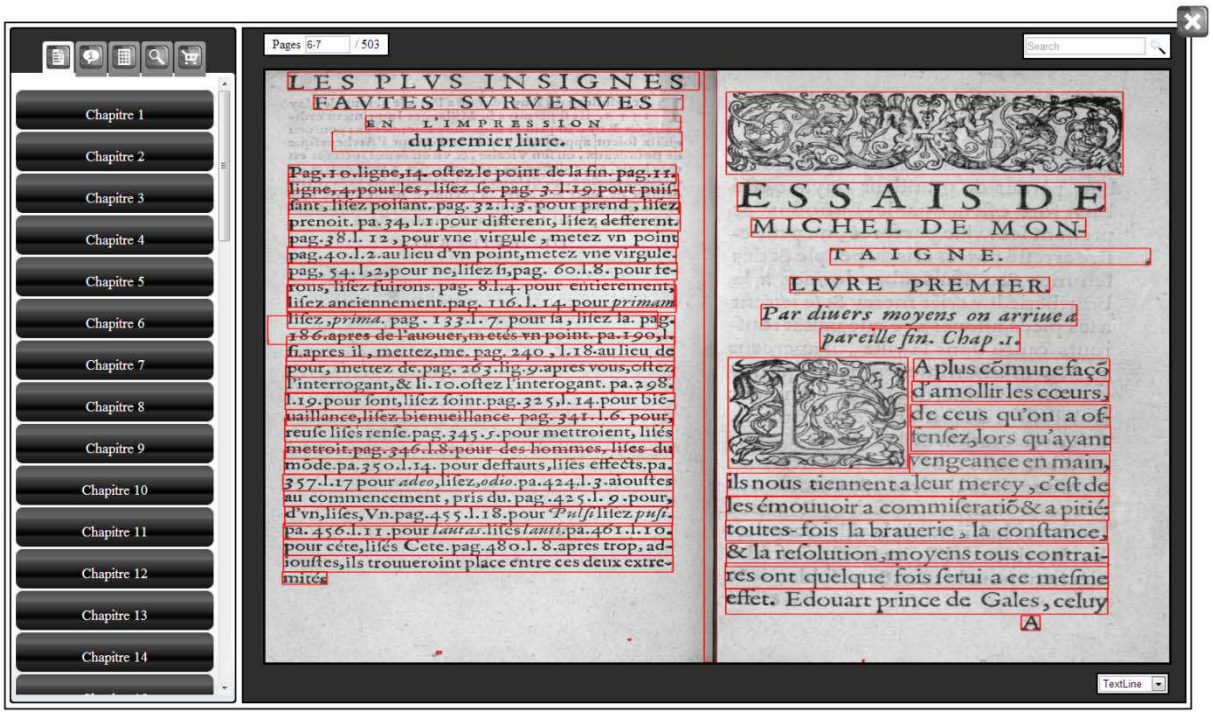

Figure 3. Screenshot of our Online Book Explorer Prototype

Acknowledgments

The work presented here was supported by the Google Digital Humanities Research Awards received by the Computer Science Laboratory of Tours - RFAI team (2010-2012).

References

[1] CESR, http://cesr.univ-tours.fr/

[2] PaRADIIT Project, https://sites.google.com/site/paradiitproject/

[3] Jean-Yves Ramel, Nicolas Sidere. Interactive Indexation and Transcription of Historical Printed Books. Digital Humanities 2011 (DH2011): June 19-22, Stanford University (USA).

[4] BVH, http://www.bvh.univ-tours.fr/presentation_en.asp

[5] See, for example, the BVH portrait gallery, http://www.bvh.univ-tours.fr/img_portrait.asp

[6a] Hendrik D. L. Vervliet: The Palaeotypography of the French Renaissance (2 vol.), BRILL, 2008

[6b] Hendrik D.L. Vervliet: French Renaissance Printing Types, The bibliographical society. OAK KNOLL Press. 2010.
[7] MUFI, http://www.mufi.info/
[8] Jacques André, “The Cassetin Project - Towards an Inventory of Ancient Types and the Related Standardised Encoding”. Proceedings of EuroTeX'2003, volume 24, No. 3, 2003, p. 314-318.
[9] CESR-BVH Exploratory Project, http://bvh.hypotheses.org/projets#Pica-Gieca, (2012-2013)
[10] Partha Pratim Roy, Frederic Rayar, Jean-Yves Ramel. An efficient Coarse-to-Fine Indexing Technique for Fast Text Retrieval in Historical Documents. In Document Analysis Systems (DAS), 27-29/03/2012.
[11] Online Historical Book Explorer prototype, http://www.rfai.li.univ-tours.fr/fr/demo/Paradiit/